SHADE

Correlation of Efficiency in Mathematics and Efficiency in other Subjects. A Statistical Study

Mathematics

A. B.

1907

UNIVERSITY OF ILL. LIBRARY

UNIVERSITY OF ILLINOIS
LIBRARY

Class 1907 Book Sh1 Volume

Je 07-10M

CORRELATION OF EFFICIENCY IN MATHEMATICS AND EFFICIENCY IN OTHER SUBJECTS.-- A STATISTICAL STUDY

BY

IMOGENE SHADE

THESIS

FOR THE

DEGREE OF BACHELOR OF ARTS

IN

MATHEMATICS

COLLEGE OF LITERATURE AND ARTS

UNIVERSITY OF ILLINOIS

PRESENTED JUNE, 1907

UNIVERSITY OF ILLINOIS

June 1 1907

THIS IS TO CERTIFY THAT THE THESIS PREPARED UNDER MY SUPERVISION BY

Imogene Shade

ENTITLED Correlation of Efficiency in Mathematics and Efficiency in Other Subjects. A Statistical Study.

IS APPROVED BY ME AS FULFILLING THIS PART OF THE REQUIREMENTS FOR THE

DEGREE OF Bachelor of Arts.

H. L. Rietz

Instructor in Charge.

APPROVED: E J Townsend

HEAD OF DEPARTMENT OF Mathematics

102013

Digitized by the Internet Archive
in 2013

http://archive.org/details/correlationofeff00shad

1907
Sh1

Correlation of Efficiency in Mathematics and Efficiency in Other Subjects – A Statistical Study.

Introduction.

A question as to the comparative efficiency of students in various lines of work has often been discussed in educational circles. It is the popular belief that every student has his own particular specialty and it is not expected that he do more than average work in branches outside of it. It is considered as usual for good work in mathematics to be accompanied by only passable work in other subjects such as history, languages and natural sciences. Of all subjects considered in adjusting courses of study the position of mathematics and the relation of the efficiency of students in this line to their efficiency in other lines has been one of the most disputed. With a view of ascertaining some definite idea of this relation we have taken up the problem of determining the correlation (1) between the efficiency of students in mathematics and in foreign languages and (2) between the efficiency of students

UNC

in mathematics and in natural sciences. Putting the problem in the form of a question it is: (1) do efficiency in mathematics and efficiency in foreign languages go together in general? and (2) do efficiency in mathematics and efficiency in natural sciences go together in general?

In a statistical problem of this sort the selection of the data is one of the most important things. Without a just and fair selection of data we cannot expect to have an unbiased result. The range of the data and the limitations placed on it influence the result more than anything else. we must select data that will give a result unbiased in all ways.

The data used in this investigation was obtained from the records of students in this university kept in the Registrar's office. In so far as possible only those cases where the student carried jointly the two subjects to be correlated were taken. In correlating mathematics and foreign languages only those cases where the student had carried the two subjects jointly during his freshman year were

considered and then the average for all the mathematical work for the year and the average for all foreign language work for the year were taken. in no case considering less than two grades in each. In order to eliminate as far as possible the probability of secondary training influencing the grades only those students were considered who had had previous training in either French or German in the high school. The data of some twelve hundred cases was taken from the years 1885-1903, the grades ranging from one hundred to fifty.

In correlating mathematics and natural sciences not so many restrictions could be imposed. Only the grades from one hundred to seventy-five were recorded and there were not enough instances where the student had carried his freshman mathematics and his elementary science at the same time so the data was confined to the average grade in mathematics and the average for all elementary natural sciences: viz. geology zoology, botany and chemistry appearing in one student's course. In

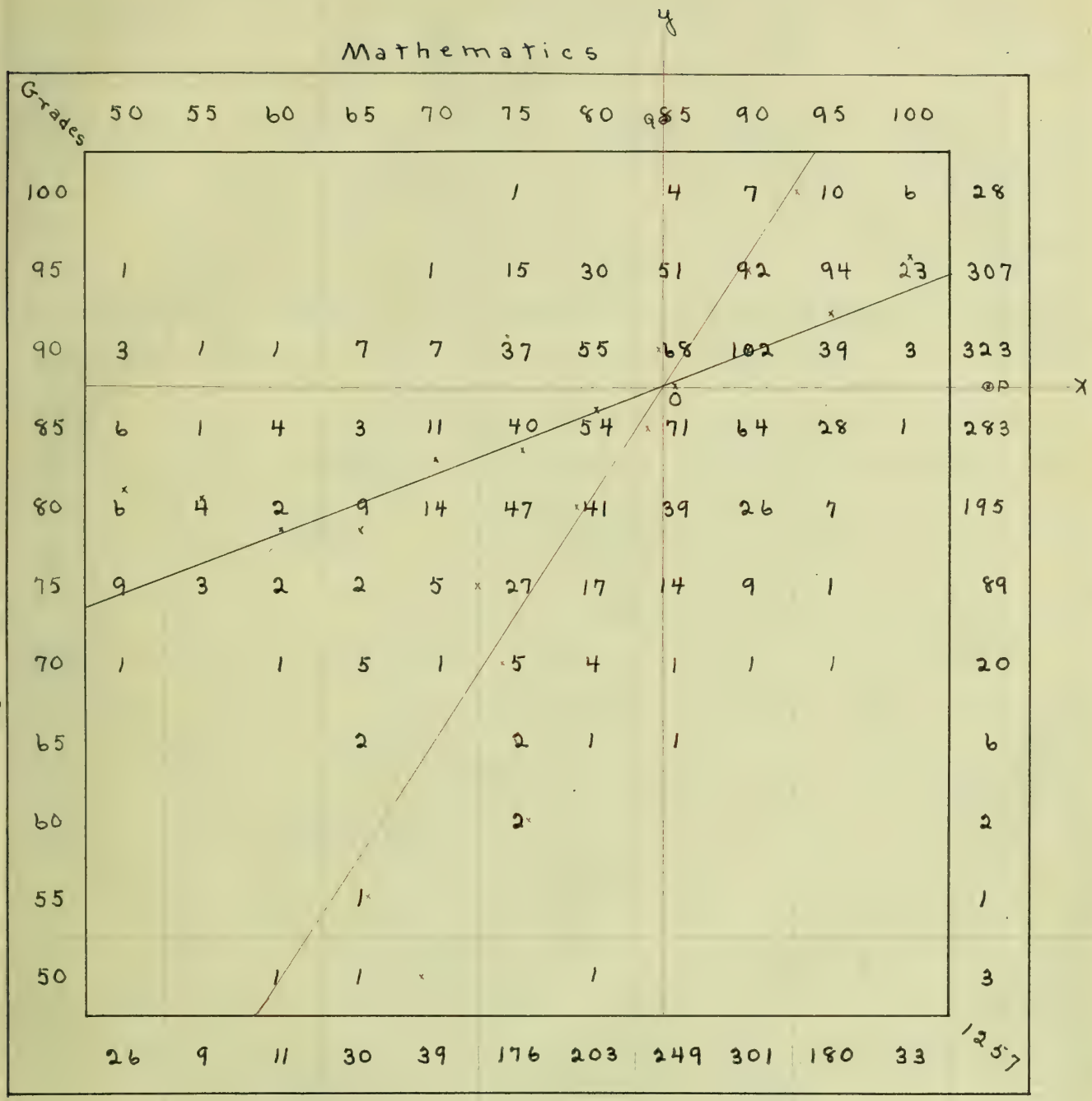
both cases no engineering students were considered because it is generally supposed that such students like mathematics before they choose their course or they wouldn't select it knowing the amount of mathematics that it contains. Now the records used come from a period in the university when mathematics was required of every one and so gives a fairer estimate of the efficiency of the average student.

The objection has been raised to collecting data thus from students' records that the result would be biased because of the student's reputation, for instance if one has the reputation of being a good student he would stand a better chance of getting a good grade than he would if his reputation wasn't known. In some few isolated cases this is probably true, but in general the instructor's judgment is not based upon the student's reputation but upon the work he does. Even the exceptional cases are more rare among the freshmen for they cannot have made much of a reputation in any line.

Tabulation

In tabulating the results, such a table as is shown in Plate I was made, the grades in mathematics at equal distances along the x axis and the grades in foreign languages at equal distances along the y axis and the whole table was thus divided into squares. In looking over the records when, for instance, a man was found who had an average of eighty-seven in mathematics and of ninety in foreign languages, a mark was placed in the square that was in the eighty-seven column and the ninety row. When a sufficient number of cases had been thus tabulated, the records in each square were counted and the number of marks inserted in the square instead of the marks themselves.

In order to lessen the amount of mathematical work the grades in both problems were tabulated in groups of five as shown in Plates II and III, i.e. the grades ranging from 88-92 inclusive were put in the ninety class or column, 93-97 inclusive in the ninety-five class



Scale: Side of melarge block
—— represents 5 percent.

Plate I

or column, etc.

In technical language these columns are known as the arrays of the ninety class, the ninety-five class, etc and the ninety, ninety-five, etc may be conveniently referred to as the "class mark".

The number of instances or individuals in any class is called the frequency of that class. A new column and a new row were made outside of the table and in these the frequencies of the classes first by rows and then by columns were tabulated, the sum of the frequency column equaling the sum of the frequency row and representing the total number of individual cases cited; in Plate I tabulated in the lower right hand square.

Modes

On looking at Plate I we see that in the column of frequencies the numbers start at twenty-eight, go up to three hundred twenty-three and then descend without any other points of maxima except at the fifty class, but this can be overlooked because when graders get as low as fifty they are irregular. The instructor

knowing that the student is below par gives him a grade without careful judgment and in general will give him a so-called "round number" like fifty, hence the apparent piling up at this point.

The point of maximum 323 at the ninety class shows us that more people get ninety as a grade in foreign language than any other grade. Thus ninety is known as the mode, the popular grade. Likewise observing the frequency row we see that here the mode is also at ninety. Ninety is the popular grade in mathematics as well as in foreign language.

Correlation Theory

Correlation according to the popular definition is merely a relation between two variable quantities. More technically speaking two characters are said to be correlated if to a selected series of sizes* of the one there correspond sizes of the other whose mean values are functions of the selected values.

The system of selected variates is

---

* The word "sizes" is here used in the sense of numerical measure

called the "subject system" and the correlated system is called the "relative system"

The main point in the correlation theory is the determination of this function mentioned in the definition.

Let $y = \phi(x)$ be the function to be determined so that $\phi(x)$ gives the mean value of the y's which correspond to a selected x.

Suppose the following system of corresponding values have been observed: (x', y') (x'', y'') ..... (x^n, y^n) where n is a positive integer large enough to give a good random sample so that a law of frequency may be assumed. These observations are said to constitute a total population or universe of observations.

As it will be more convenient in this theory to deal with deviations from the means of the correlated populations than with the observations themselves, let (x_1, y_1) (x_2, y_2) ..... (x_n, y_n) represent deviations from the mean values. We may well imagine these deviations represented with respect to coordinate axes through the means of the total population.

In a practical application such as we have in mind the means of the total populations may be found as we have represented them on Plate I by P and Q and the horizontal and vertical lines drawn through these points are the coordinate axes and their intersection is the origin of the table and at the mean value for the total population.

Now we may assume for this theory that our observations have been grouped into classes in such a way as to best show the frequency distribution.

Let s be the number of classes, and x_1' x_2' .... x_s' be the class marks so that the observed X's are now grouped into s classes.

In each array as shown in Plate I we find the class mean i.e. the average grade in foreign languages for all those students with class mark 100 in mathematics, 95 in mathematics etc. In our notation the class mean of the array of class mark x_1 is $\frac{\Sigma f y}{n_1}$ where the y's are the grades in foreign language and n_1 the number of individuals in the first

array with class mark x_1. Similarly we obtain the means of all the s arrays. The points marked × Plate I represent these means.

It is sometimes convenient to refer to the distributions in the vertical columns as the y arrays instead of calling them the arrays of class mark x_1, x_2 ... etc.

These points, the class means of the y arrays do not lie at random over the figure but tend to arrange themselves approximately along a straight line. So the function $\phi(x)$ which is to be determined is a linear function and our correlation is a case of linear correlation.

Our line along which the class means of the y arrays tend to arrange themselves is known as the line of regression of foreign languages on mathematics.

If the means lie exactly on the line, the regression is said to be "truly linear."

We could of course find the class means of each of the x arrays or arrays of

class marks $y_1, y_2 \cdots y_n$. These means would arrange themselves along another line known as the regression line of mathematics on foreign languages.

Having thus found the class means of both the y and x arrays we can find the origin of the coordinate axes by taking the mean of the class means of the y's weighted with their number and the mean of the class means of the x's weighted with their number. We draw horizontal and vertical lines through the points thus obtained, and their intersection is the origin of coordinates.

Derivation of regression line.

To obtain the line which will best fit these points we must subject it to some minimal condition. If we selected the usual minimal condition* namely that the summation of the squares of the deviations of the means of arrays from the line shall be a minimum when the deviations are measured

* Yule, Statistical Journal, Royal Society; Vol. 60 p812-851, 1897

parallel to the y axis and each deviation is weighted with the number of points in the array, we obtain a simple form for the line of regression.

Let $y = mx + b$ be the equation of the regression line, then the whole question in determining the line is to so determine m and b that the values of y calculated from them shall differ as little as possible from the means of the observed values.

The above minimal condition requires that

(A) $\sum_{t=1}^{t=s} n_t (\bar{y}_t - mx'_t - b)^2$ * shall be a minimum where n_t is the number of variates in the array of class mark x'_t

Differentiating partially with respect to m and b

$$-2 \Sigma n_t x'_t (\bar{y}_t - mx'_t - b) = 0 \qquad (1)$$

$$-2 \Sigma n_t (\bar{y}_t - mx'_t - b) = 0 \qquad (2)$$

But $\Sigma n_t \bar{y}_t = 0$ and $\Sigma n_t x'_t = 0$ since the mean of the total population is at the origin.

Therefore b must be 0

From (1) $\sum_1^s n_t \bar{y}_t x'_t - m \Sigma n_t x'^2_t = 0$

This is equivalent to

* $\bar{y}$ represents the mean value in a y array.

(B) $\sum_1^s (\Sigma y)_t x'_t - m \sum_1^s n_t x'^2_t = 0$ where $(\Sigma y)_t$ is the sum of the observed y's in the array of mark x'_t

When the population n approaches infinity the class marks and the observed x's may be taken as equal. Hence (B) becomes

(C) $\sum_1^n y_d x_d - m \sum_1^n x_d^2 = 0$ where the summation on the d extends to corresponding observed deviations from the means.

At this point the work can be thrown into a more convenient form if we put

$$\sigma_y^2 = \frac{y_1^2 + y_2^2 + \cdots + y_n^2}{n} = \frac{\sum_1^n y_d^2}{n} \quad (1)$$

and $$\sigma_x^2 = \frac{x_1^2 + x_2^2 \cdots + x_n^2}{n} = \frac{\sum_1^n x_d^2}{n} \quad (2)$$

When σ_y and σ_x are known as the standard deviations of the y-population and the x-population respectively. The standard deviations are the same as the "mean error" of Gauss.

So from (C) using the values in (1) and (2) we have

$$m = \frac{\sum_1^n y_d x_d}{n \sigma_x^2} = \frac{\Sigma x_d y_d}{n \sigma_x \sigma_y} \frac{\sigma_y}{\sigma_x}$$

So we have for the equation of our line

$$y = \frac{\Sigma x_d y_d}{n \sigma_x \sigma_y} \frac{\sigma_y}{\sigma_x} x \quad \text{or,}$$

(D) $y = r\frac{\sigma_y}{\sigma_x}x$ where r equals $\frac{\Sigma x_d y_d}{n\sigma_x\sigma_y}$.

This is the equation of the regression line which was to be derived. It is called the regression equation of y with respect to x. For a selected x the expression $r\frac{\sigma_y}{\sigma_x}$ gives a value around which the observed values of y center.

If the regression is truly linear this expression gives exactly the mean value of the y's in an array; if it is only approximately linear it gives the approximate mean value.

By parity of reasoning the regression equation of x with respect to y is

(E) $x = r\frac{\sigma_x}{\sigma_y}y$

It should be observed that one cannot solve the regression equation (D) for the value of x in (E). The correspondence in these equations is not a "one to one" correspondence in the usual sense, it is a correspondence between selected values in the means.

Correlation coefficient

The correlation coefficient r which is equal to $\frac{\sum_1^n x_d y_d}{n\sigma_x\sigma_y}$ is not only a good measure for correlation but is proved

by Pearson to be the best* when the distribution is normal.

It is of importance to note that this derivation makes no assumptions as to the nature of the frequency distribution while all the derivations preceding that of Yule assumed a normal law of distribution†. The significance of the correlation coefficient does not necessarily depend upon the distribution being normal but rather upon the fact that the means of the arrays would not be far from the line of regression.

It may be worth remarking that a test as to how nearly the distributions of mathematics and foreign languages are normal, showed that the discrepancies were such that it could at best be said to fit this curve roughly and that we therefore rest our argument on no such assumption.

---

* Philosophical Transactions Royal Society Vol 187A 1896 p 264-

† Proceedings of the Royal Society Vol 60 p. 477-489

Regression coefficient.

The term regression is one borrowed from evolution where individuals are said to regress toward the mean of the race. The coefficient $r\frac{\sigma_r}{\sigma_s}$ where the relative system is correlated with the subject system is large where there is little regression and may well be popularly called the "coefficient of prediction". The expression $r\frac{\sigma_y}{\sigma_x}x$ gives the ordinates of the line of regression; $r\frac{\sigma_y}{\sigma_x}$ being the coefficient of regression and often denoted by R.

Limits for r.

Taking the sum of the squares of the differences between the observed y's and the values obtained for the y's by substituting the corresponding x's in $r\frac{\sigma_y}{\sigma_x}x$ we have in algebraic form:

$$\sum_1^m{}_p\left(y_p - r\frac{\sigma_y}{\sigma_x}x_p\right)^2 = \Sigma y_p^2 - 2r\frac{\sigma_y}{\sigma_x}\Sigma y_p x_p + r^2\frac{\sigma_y^2}{\sigma_x^2}\Sigma x_p^2$$

$$= n\sigma_y^2 - 2r\frac{\sigma_y}{\sigma_x}n\sigma_y\sigma_x + r^2\frac{\sigma_y^2}{\sigma_x^2}n\sigma_x^2$$

(F) $$\sum_1^m{}_p\left(y_p - r\frac{\sigma_y}{\sigma_x}x_p\right)^2 = n\sigma_y^2(1-r^2)$$

From (F) r must lie between +1 and −1.

When $r=1$ the correlation is perfect i.e. all the points lie on the regression line. In other words when one character is given, the corresponding value of the other could

be exactly given for each individual case.

When $r = -1$ we have perfect negative correlation i.e. the line would lie at right angles to the position it has when $r = +1$.

When $r = 0$ all the mean values of the y arrays would fall on the x axis under the independent action of the laws of probability.

On Plate I we have shown both of the regression lines i.e. the regression of foreign languages on mathematics and the regression of mathematics on foreign languages. The second was found similarly to the first, namely by finding the mean values of the x arrays and passing a straight line among them subject to the above minimal condition.

The regression coefficient is the slope of the regression line and in order to get the slope of the mathematics on foreign language line we must consider the tangent of the angle between this line and the y axis instead of the x axis as we take it in the other case.

The guess method.

In dealing with the large numbers which we have in these problems, it seemed practical to adopt some method which would somewhat simplify the mathematical calculations, so we have adopted a method which Yule has suggested.*

Referring to Plate II we have first our column of frequencies of foreign language with respect to mathematics, then instead of actually finding at the start what the mean grade of foreign language in we make a guess, the value represented by g, in our case 85, and make a column of $(v-g)$'s where v stands for the true grade received at each class mark y_1, $y_2 \cdots y_m$ respectfully i.e. the $(v-g)$'s have the values $(100-85), (95-85), (90-85)$, etc. Then we multiply the f column into the $(v-g)$ column thus making a column $f(v-g)$. $\frac{\Sigma f(v-g)}{n}$ gives the average deviation from the assumed axis of foreign language, eighty-five.

$$\frac{\Sigma f(v-g)}{n} = 2.096 \text{ by Plate II}$$

Hence we have 85 + 2.096 or 87.096 as the

* Statistical Journal, Royal Society Vol 60 pp812-851

Foreign Languages

	50	55	60	65	70	75	80	85	90	95	100	f	v-q	f(v-q)	(v-q)²	f(v-q)²	D'M D'L
100						1		4	7	10	6	28	15	420	225	6300	3225
95	1				1	15	30	51	92	94	23	307	10	3070	100	30700	13950
90	3	1	1	7	7	37	55	68	102	39	3	323	5	1615	25	8075	-525
85	6	1	4	3	11	40	54	71	64	28	1	283					
80	6	4	2	9	14	47	41	39	26	7		195	-5	-975	25	4875	6225
75	9	3	2	2	5	27	17	14	9	1		89	-10	-890	100	8900	8700
70	1		1	5	1	5	4	1	1	1		20	-15	-300	225	4500	3450
65				2		2	1	1				6	-20	-120	400	2400	1300
60						2						2	-25	-50	625	1250	500
55				1								1	-30	-30	900	900	600
50			1	1			1					3	-35	-105	1225	3675	1750
f	26	9	11	30	39	176	203	249	301	180	33	1257		2635		71575	39175
v-q	-35	-30	-25	-20	-15	-10	-5		5	10	15			2.096		56.941	
f(v-q)²	1225	900	625	400	225	100	25		25	100	225						
f(v-q)	-910	-270	-275	-600	-585	-1760	-1015		1505	1800	495			-1615			
D'M f(v-q)²	31800	8100	6875	12000	8775	17600	5075		7525	18000	7425			123175			97.991
D'L D'M	3850	1350	1875	4200	1350	3200	-425		6500	12250	5025			39175			
														31.165			

$\sigma_L = \sqrt{56.941 - \overline{2.096}^2} = 7.25$

$\sigma_M = \sqrt{97.991 - \overline{1.285}^2} = 9.82$

$\sigma_L \sigma_M = 71.195$

$\frac{\Sigma D_M' D_L'}{n} = \frac{39175}{1257} = 31.165$

$c_L c_M = (2.096)(-1.285) = -2.693$

$r = \left\{\frac{\Sigma D_M' D_L'}{n} - c_L c_M\right\} \frac{1}{\sigma_L \sigma_M} = (31.165 + 2.693) \div 71.195 = .476$

$R = \frac{\sigma_L}{\sigma_M} r = \frac{7.25}{9.82} .476 = .351$

$E_r = .67449 \frac{1 - r^2}{\sqrt{n}} = .67449 \frac{.773424}{35.45} = .0147$

$E_R = .67449 \frac{\sigma_L}{\sigma_M} \sqrt{\frac{1 - r^2}{n}} = .67449 \times 0.738 \times 0.0248 = .0123$

Plate II

average or mean value of grades in foreign language. This value corresponds with that obtained for the axis of means for foreign language in Plate I.

$\frac{\Sigma f(v-g)}{n}$ is known as the correction, the c_L in Plate II. It is the amount which added to the assumed axis gives the true axis.

Next we form a column of the squares of the deviations from the guess, $(v-g)^2$'s.

$\frac{\Sigma f(v-g)^2}{n}$ gives the average square of the deviation from the assumed axis.

Derivation of σ

From ① page 14 we see that σ^2 equals $\frac{\Sigma y^2}{n}$

We have now to show that

$\frac{\Sigma y^2}{n} = \frac{\Sigma f(v-g)^2}{n} - \left(\frac{\Sigma f(v-g)}{n}\right)^2$ in our notation shown in Plate II.

$y = V - M$ where M is the mean value

$M = g + c$ where g is the guess and c is equal to $\frac{\Sigma f(v-g)}{n}$.

$$\frac{\Sigma y^2}{n} = \frac{\Sigma f(v-M)^2}{n} = \frac{\Sigma f(v-g-c)^2}{n} = \frac{\Sigma f(v-g)^2}{n} - \frac{2c\,\Sigma f(v-g)}{n} +$$

$$\frac{nc^2}{n} = \frac{\Sigma f(v-g)^2}{n} - 2c\,\frac{\Sigma f(v-g)}{n} + c^2 = \frac{\Sigma f(v-g)^2}{n} - c^2$$

Since c^2 equals $\left(\frac{\Sigma f(v-g)}{n}\right)^2$ we have

$$\frac{\Sigma y^2}{n} = \frac{\Sigma f(v-g)^2}{n} - \left(\frac{\Sigma f(v-g)}{n}\right)^2 \quad \text{Q. E. D.}$$

Applying this formula to the data on Plate II we find the standard deviation of

foreign languages to be 7.25

Now just as we have made the columns f, $(v-g)$, $f(v-g)$, $(v-g)^2$ and $f(v-g)^2$ for foreign languages we make similar rows for mathematics. We find the average deviation from the assumed mean of mathematics to be -1.2848. So our axis of means for mathematics is 85 - 1.2848 or 83.7552.

The standard deviation of mathematics, σ_M is 9.82

Derivation of Correlation coefficient.

We make a column and a row of product deviations designated by $D'_M D'_L$ on Plate II i.e we find for each variate the product of its deviations from the respective assumed means. In the first instance we have $[1(-10) + 7(5) + 10(10) + 6(15)]\,15 = 3225$. With due regard to sign we take the algebraic sum of these products.

On page 15 we saw that the correlation coefficient r was equal to $\frac{\sum x_d y_d}{n \sigma_x \sigma_y}$, we have now to prove that

$$\frac{\Sigma x_d y_d}{n \sigma_x \sigma_y} = \left\{ \frac{\Sigma D'_M D'_L}{n} - c_m c_L \right\} \frac{1}{\sigma_M \sigma_L}$$ where $x_d y_d$ are

the product deviations from the true mean, $D'_M D'_L$ the product deviations from the

assumed mean and σ_x and σ_y are the same as σ_M and σ_L in our new notation.

$$\left.\begin{aligned} M_M &= G_M + c_M \\ M_L &= G_L + c_L \end{aligned}\right\} (a) \qquad \left.\begin{aligned} x_d &= V_M - M_M \\ y_d &= V_L - M_L \end{aligned}\right\} (b) \qquad \left.\begin{aligned} D'_M &= V_M - G_M \\ D'_L &= V_L - G_L \end{aligned}\right\} (c)$$

$$\frac{\Sigma x_d y_d}{n \sigma_x \sigma_y} = \frac{\Sigma (V_M - M_M)(V_L - M_L)}{n \sigma_M \sigma_L} = \frac{\Sigma (V_M - G_M - c_M)(V_L - G_L - c_L)}{n \sigma_M \sigma_L} =$$

$$\frac{\Sigma (D'_M - c_M)(D'_L - c_L)}{n \sigma_M \sigma_L} = \frac{\Sigma D'_M D'_L - c_L \Sigma D'_M - c_M \Sigma D'_L + \Sigma c_M c_L}{n \sigma_M \sigma_L} =$$

$$\frac{\Sigma D'_M D'_L - c_L \Sigma (V_M - G_M) - c_M \Sigma (V_L - G_L) + \Sigma c_M c_L}{n \sigma_M \sigma_L} =$$

$$\frac{\Sigma D'_M D'_L - c_L \Sigma (V_M - M_M + c_M) - c_M \Sigma (V_L - M_L + c_L) + \Sigma c_M c_L}{n \sigma_M \sigma_L}$$

But here $\Sigma(V_M - M_M + c_M)$ and $\Sigma (V_L - M_L + c_L)$ are nothing more than $n c_M$ and $n c_L$, so we have

$$\frac{\Sigma x_d y_d}{n \sigma_x \sigma_y} = \frac{\Sigma D'_M D'_L - (\Sigma c_M c_L)}{n \sigma_M \sigma_L} = \left\{ \frac{\Sigma D'_M D'_L}{n} - \frac{n c_M c_L}{n} \right\} \frac{1}{\sigma_M \sigma_L}$$

$$r_{xy} = \left\{ \frac{\Sigma D'_M D'_L}{n} - c_M c_L \right\} \frac{1}{\sigma_M \sigma_L} \qquad \text{Q. E. D.}$$

Substituting the values in this formula we get .476 as a correlation coefficient.

The regression coefficient equal to $r \frac{\sigma_L}{\sigma_M}$ or $.476 \frac{7.25}{9.82}$ is .351. This is numerically the slope of the regression line, in other words the numerical value of the tangent of the angle between the line of regression

of foreign language on mathematics and the x axis

Probable Error

The probable error in both r and R may be calculated from the formulae –

$$E_r = .67449 \frac{1-r^2}{\sqrt{n}} \quad \text{and} \quad E_R = .67449 \frac{\sigma_1}{\sigma_0}\sqrt{\frac{1-r^2}{n}} \; *$$

It may be worth while discussing briefly the significance of probable error in any result such as we have obtained. It is a measure designed to indicate what degree of confidence is to be placed in a statistical determination.

The "true value" in statistical work is difficult of accurate definition and cannot be actually determined. It may however be thought of as that ideal value which would result from dealing with an infinite population of which the given finite population (in our case 1257) is a random sample.

The probable error in any result may be defined as that deviation from the determined value on either side such that it is an even wager that the true

---

* Pearson and Filon, Proc. Roy. Soc. 1898 page 242.

value lies within this amount of the determined value i.e. the true value of r lies someplace between .476 + .0147 and .476 - .0147.

Correlation of Mathematics and Natural Sciences.

Plate III. IV and V show the correlation of mathematics with natural sciences. From Plate III we see that the mode i.e. the popular grade in natural science is eighty-five and the mode in mathematics is also eighty-five. The means lie at 86.2 and 85.03 for natural science and mathematics respectively.

The regression line of natural science on mathematics has the equation $y = .438x$, in other words the regression coefficient, the slope of the line is .438. The regression line of mathematics on natural science, the red line in Plate III has the equation $x = .44y$. in other words the tangent of the angle which this line makes with the y axis has the value .44.

From Plate IV we see that the standard deviation of natural science is 6.67 and the standard deviation of math-

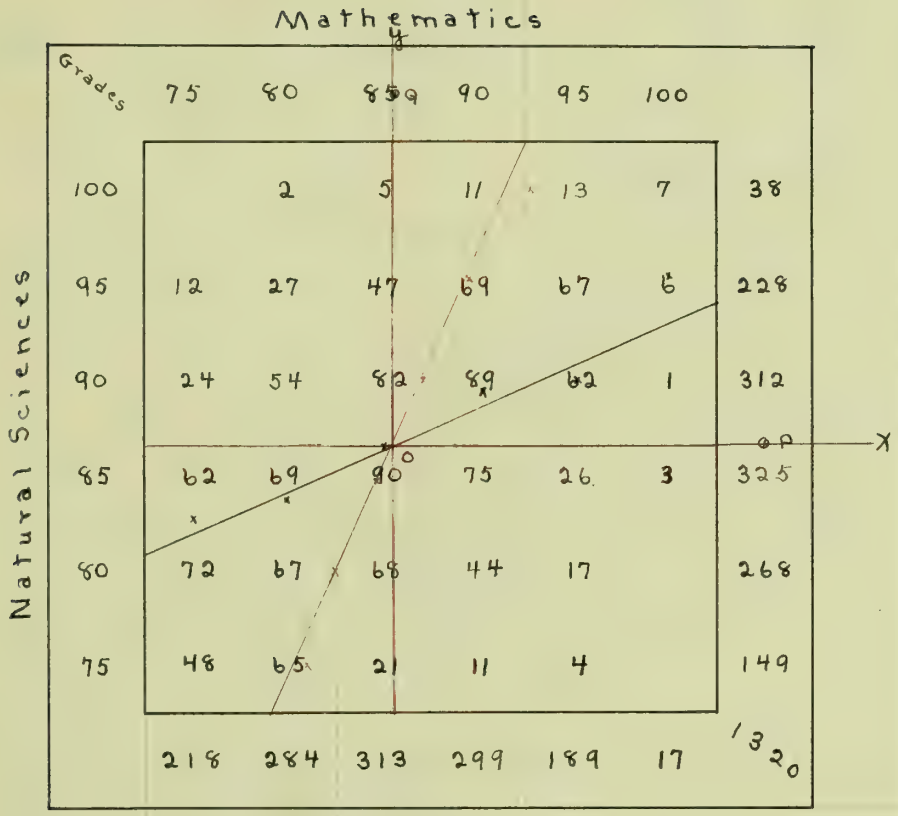

Mathematics
Natural Sciences
Grades
75 80 85 90 95 100
100 2 5 11 13 7 38
95 12 27 47 69 67 6 228
90 24 54 82 89 62 1 312
85 62 69 90 75 26 3 325
80 72 67 68 44 17 268
75 48 65 21 11 4 149
218 284 313 299 189 17 1320
Y
X

Plate III

Natural Sciences

Grades	75	80	85	90	95	100	f	v-g	f(v-g)	$(v-g)^2$	$f(v-g)^2$	$D_M' D_N'$
100		2	5	11	13	7	38	15	570	225	8550	4200
95	12	27	47	69	67	6	228	10	2280	100	22800	8500
90	24	54	82	89	62	1	312	5	1560	25	7800	2850
85	62	69	90	75	26	3	325					
80	72	67	68	44	17		268	-5	-1340	25	6700	3325
75	48	65	21	11	4		149	-10	-1490	100	14900	7100
f	216	264	313	299	189	17	1320		1580 / 1320 = 1.197		60750 / 1320 = 46.023	25975 / 1320 = 19.68
(v-g)	-10	-5		5	10	15						
f(v-g)	-2160	-1420		1495	1890	255	40 / 1320 = .0303					
$(v-g)^2$	100	25		25	100	225						
$f(v-g)^2$	21600	7100		7475	18900	3825	59100 / 1320 = 44.773					
$D_N' D_M'$	6000	2075		4850	10500	2550	25975 / 1320 = 19.68					

$\sigma_N^2 = 46.023 - \overline{1.197}^2 = 6.677$

$\sigma_M^2 = 44.77 - \overline{.0303}^2 = 6.691$

$\sigma_N \cdot \sigma_M = 44.676$

$\frac{\Sigma D_M' D_N'}{n} = 19.68$

$C_M C_N = .0303 \times 1.197 = .036$

$r = \left\{ \frac{\Sigma D_M' D_N'}{n} - C_M C_N \right\} \frac{1}{\sigma_M \sigma_N} = .439$

$R = \frac{\sigma_N}{\sigma_M} r = .438$

$E_r = .67449 \frac{1-r^2}{\sqrt{n}} = .0149$

$E_R = .67449 \frac{\sigma_N}{\sigma_M} \sqrt{\frac{1-r^2}{n}} = .0166$

Plate IV

ematics from the mean is 6.69

The correlation coefficient in this instance was .439, not quite so large as in the case of mathematics and languages.

The probable error in r is .0149 and in R is .0166

On Plate V we have tabulated the data for mathematics and natural sciences in groups of one in order to compare the results of this method of grouping with those results obtained by grouping in fives as shown in Plate IV.

The results come out approximately the same and the amount of numerical work in Plate IV is much less than that involved in Plate V while the former has the advantage of showing more distinctly in what region the mode lies. The frequency columns in Plate V show so many points of maxima that it is impossible to tell which is the popular grade.

As for the significance of the correlation coefficient in these two cases the difference is only .009 and as the probable error includes this, the coefficient

Statistics (Yadura)

	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	f	(v-9)	f(v-9)	(v-9)²	f(v-9)²	$D'_M D'_N$
100															2	2				1							5	15	75	225	1225	405
99																3			1			1	1	2	1	1	10	14	140	196	1960	1414
98						1		1	1	1	3					3		1	2	3	3		1	2		1	23	13	299	169	3887	1781
97	1			1		3	1	1					1	1	3	1	1	1			5	2	2	1			25	12	300	144	3600	1260
96						1	2	3	2		1	1	2	2	2	2	1	3	2	2	2	7			1		36	11	396	121	4356	1925
95			1			2	1	1	3	1	5	1	6		2	8	6	7	3	3	8	6	2	3	1		70	10	700	100	7000	3880
94	2		1			2		1	1	1	3	3	1		1	3	1	1	3	4	2	2	1				33	9	297	81	2673	963
93	6		1	3	1	2		2		5	5	1	4	4	2	7	6	4	2	1	5	3					64	8	512	64	4096	960
92	1	2	3			4	1	1	2	1	5		5		3	4	2	4	2	3	1	1	2				47	7	329	49	2303	602
91			1	2	2	4		1		4	3	4	5		4	3	7	4	4	1	2	1	1				53	6	318	36	1908	828
90	8		4		2	7	5	7	2	4	9	5	8	3	4	8	6	3	7	5	13	6	2	1			119	5	595	25	2975	1345
89	3			2	1	3		2	1	2	5	2	2	1	1	2	8	6	1	2	3						47	4	188	16	752	376
88	2				2	6	1	1		4	3	2	4	2	2	4	4	4	1	1	1		2				46	3	138	9	414	222
87	3	2	2	1		2	2	1	1	3	4		5	3	7	3	5	2	1	3	4		1	1		1	57	2	114	4	228	248
86	5	3	2	2	2	1	1	1		1	1	4	2	2	3	1			1								32	1	32	1	32	-93
85	18	1	13	2	6	9	5	7	3	9	13	3	17	5	6	10	8	2	3	2	3	1	1	1			148					
84	4		2	2	2	3	2	3	2	2	5	3	3		4	1	3	4	2		1						48	-1	-48	1	48	18
83	4		3	2	4	6	2	1	2		5		2	1		1	1	3	1	1	1						40	-2	-80	4	160	162
82	11	1	9	2	3	7		2	2	1	4	3	5	5	2	1	2	3			2	1					66	-3	-198	9	594	492
81	4	3	4	1	2		1	3	2	2	3	2	1	3		1	2					1					35	-4	-140	16	560	384
80	12	7	12	3	5	13	6	9	5	4	11	3	11	7	3	5	4	2	3	2	2	3	2				134	-5	-670	25	3350	1100
79	4		1	1	1					1		1	2	1	1												13	-6	-78	36	468	300
78	3		1	1	2	3	1	1		1	2		2	1			1		1								20	-7	-140	49	980	413
77	8		5	3	1	1	5	1		1	1		1														27	-8	-216	64	1728	1392
76	3		1	1	1	2		1													1						10	-9	-90	81	810	486
75	16	3	12	10	4	20	4	11	4	2	6	1	5	5	2	2	1	1	1		1		1				112	-10	-1120	100	11200	4490
f	118	22	78	39	41	102	40	62	33	50	97	39	94	46	54	75	69	55	41	34	60	35	19	11	3	3	Σf_0		1653		57307	25353
																													1.252		43.414	
(v-9)	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15						
f(v-9)	-1180	-198	-624	-273	-246	-510	-160	-186	-66	-50		39	188	138	216	375	414	385	328	306	600	385	228	143	42	45	2339					
																											2339					
																											2082					
																											.257					
$(v-9)^2$	100	81	64	49	36	25	16	9	4	1		1	4	9	16	25	36	49	64	81	100	121	144	169	196	225						
$f(v-9)^2$	11800	1782	4992	1911	1476	2550	640	558	132	50		39	376	414	864	1875	2484	2695	2624	2754	6000	4235	2736	1859	588	675	36109					
																											42.506					
$D'_M D'_N$	2630	531	1704	819	522	560	172	141	-38	-76		72	214	-84	696	1830	1344	1659	1456	1863	3280	2618	1176	1339	490	435	25353					
																											25353					

$$\sigma_N = 43.414 - \overline{1.252}^2 = 6.468$$

$$\sigma_M = 42.506 - \overline{.257}^2 = 6.514$$

$$\sigma_N \sigma_M = 42.132$$

$$\frac{\Sigma D'_M D'_N}{n} = 19.206$$

$$c_N c_M = 1.252 \times .257 = .3215$$

$$r = \left\{ \frac{\Sigma D'_M D'_N}{n} - c_N c_M \right\} \frac{1}{\sigma_N \sigma_M} = .448$$

$$R = \frac{\sigma_N}{\sigma_M} r = \frac{6.468}{6.514} .448 = .445$$

$$E_r = .67449 \frac{1-r^2}{\sqrt{n}} = .0148$$

$$E_R = .67449 \frac{\sigma_N}{\sigma_M} \sqrt{\frac{1-r^2}{n}} = .0165$$

Plate V

obtained in the first instance is just as correct a measure of the correlation as that obtained in the second case.

General Conclusions

Considering the investigation as a whole we see that the correlation between mathematics and languages is somewhat closer than that between mathematics and natural sciences. At first this may seem very strange and it is a conclusion which doesn't correspond with the popular notion that mathematics being a science is more closely correlated with other sciences than with language. However a considerable portion of the difficulty which students have with mathematics arises from a failure to understand the symbolism or language in which it is written. The ability required to translate a thought from an English sentence into algebraic language is the same kind of ability required to translate English into a foreign language. So efficiency in foreign language would have a direct bearing upon efficiency in this phase of math-

matical work. Furthermore the comparatively loose inductive methods employed in natural sciences would not prepare one for the strictly logical arguments of pure mathematics.

The popular grade in mathematics was ninety in the first correlation but eighty-five in the second. This apparent fallacy is explained by the fact that the grades used in the second correlation were taken from a later period in the university when a closer system of marking was used

It may be said in connection with the theory of formal discipline that our results substantiate the theory in a narrow sense but have no bearing upon the belief that efficiency in one line will make a person efficient in every department of life. It does show that so far as subjects considered in this investigation are concerned that power in mathematics is accompanied by power in language and in natural sciences.

There is no question as to the significance of our results. There is a

very close correlation in both cases which means that the average student who is good in mathematics is also good in foreign languages and in natural sciences.

The probable error in both cases is so small in comparison with the coefficients themselves that it is safe to say that our results are close to the true values. If we should allow a deviation from our determined value of five times the probable error the wager would stand about thirteen hundred to one that our result differs from the true value by no more than this amount.

That the reader may appreciate the amount of correlation that really does exist between the subjects used in our investigation we shall give a few of the well known correlation coefficients for closely related groups of individuals. We have confined ourselves to the correlation coefficients pertaining to stature as they are more readily understood.

We have the three following

coefficients:*

(1) Stature of father and son .396

(2) Stature of brother and brother .391

(3) Stature of sister and sister .444

The first coefficient is known among evolutionists as one of high value. Both of our coefficients are higher than this value and one is even higher than the third. viz; the correlation coefficient of mathematics and foreign languages.

So our correlation coefficients which we have obtained may be considered as high values and in conclusion we say as a result of this investigation that efficiency in mathematics and efficiency in foreign languages go together in general and in very nearly the same degree do efficiency in mathematics and efficiency in natural sciences go together.

* Grammar of Science - Karl Pearson page 458

UNIVERSITY OF ILLINOIS-URBANA
3 0112 086764153